Data Grids, Digital Libraries, and Persistent Archives

(Storage Resource Broker - SRB)

Arcot Rajasekar
Michael Wan
Reagan W. Moore
(sekar, mwan, moore)@sdsc.edu

Topics

- Concepts behind data management
- Production data grid examples
- Integration of data grids with digital libraries and persistent archives

Data Management Concepts (Elements)

- Collection
  - The organization of digital entities to simplify management and access.
- Context
  - The information that describes the digital entities in a collection.
- Content
  - The digital entities in a collection

Types of Context Metadata

- Descriptive
  - Provenance information, discovery attributes
- Administrative
  - Location, ownership, size, time stamps
- Structural
  - Data model, internal components
- Behavioral
  - Display and manipulation operations
- Authenticity
  - Audit trails, checksums, access controls

Metadata Standards

- METS - Metadata Encoding Transmission Standard
  - Defines standard structure and schema extension
- OAIS - Open Archival Information System
  - Preservation packages for submission, archiving, distribution
- OAI - Open Archives Initiative
  - Metadata retrieval based on Dublin Core provenance attributes

Data Management Concepts (Mechanisms)

- Curation
  - The process of creating the context
- Closure
  - Assertion that the collection has global properties, including completeness and homogeneity under specified operations
- Consistency
  - Assertion that the context represents the content

Information Technologies

- Data collecting
  - **Sensor systems**, object ring buffers and portals
- Data organization
  - **Collections**, manage data context
- Data sharing
  - **Data grids**, manage heterogeneity
- Data publication
  - **Digital libraries**, support discovery
- Data preservation
  - **Persistent archives**, manage technology evolution
- Data analysis
  - **Processing pipelines**, manage knowledge extraction

Data Management Challenges

- Distributed data sources
  - Management across administrative domains
- Heterogeneity
  - Multiple types of storage repositories
- Scalability
  - Support for billions of digital entities, PBs of data
- Preservation
  - Management of technology evolution

Data Grids

- Distributed data sources
  - Inter-realm authentication and authorization
- Heterogeneity
  - Storage repository abstraction
- Scalability
  - Differentiation between context and content management
- Preservation
  - Support for automated processing (migration, archival processes)

Assertion

- Data Grids provide the underlying abstractions required to support
  - Digital libraries
    - Curation processes
    - Distributed collections
    - Discovery and presentation services
  - Persistent archives
    - Management of technology evolution
    - Preservation of authenticity

SRB Collections at SDSC

	As of 12/22/2000		As of 5/17/2002		As of 11/14/2003		
Project Instance	*Data_size (in GB)*	*Count (files)*	*Data_size (in GB)*	*Count (files)*	*Data_size (in GB)*	*Count (files)*	*Users*
Data Grid							
Digsky	7,599.00	3,630,300	17,800.00	5,139,249	42,786.00	6,076,982	69
NPACI	329.63	46,844	1,972.00	1,083,230	8,822.00	2,995,432	377
Hayden			6,800.00	41,391	7,835.00	60,001	168
SLAC			514.00	77,168	2,108.00	294,149	43
LDAS/SALK			239.00	1,766	824.00	13,016	66
TeraGrid					10,603.00	433,938	2,229
BIRN					389.00	1,084,749	167
Digital Library							
DigEmbryo	124.30	2,479	433.00	31,629	720.00	45,365	23
HyperLter	28.94	69	158.00	3,596	215.00	5,097	28
Portal			33.00	5,485	1,244.00	34,094	352
AfCS			27.00	4,007	107.00	21,295	21
NSDL/SIO Exp			19.20	383	603.00	87,191	26
TRA			5.80	92	92.00	2,387	26
SCEC					12,274.00	1,721,241	43
UCSDLib					1,085.00	138,421	29
Persistent Archive							
NARA/Collection			7.00	2,455	67.00	82,031	56
NSDL/CI					465.00	2,948,903	114
TOTAL	8 TB	3.7 million	28 TB	6.4 million	90 TB	16 million	3837

** Does not cover data brokered by SRB spaces administered outside SDSC.
Does not cover databases; covers only files stored in file systems and archival storage systems
Does not cover shadow-linked directories

Common Infrastructure

- Digital libraries and persistent archives can be built on data grids
- Common capabilities are needed for each environment
- Multiple examples of production systems across scientific disciplines and federal agencies

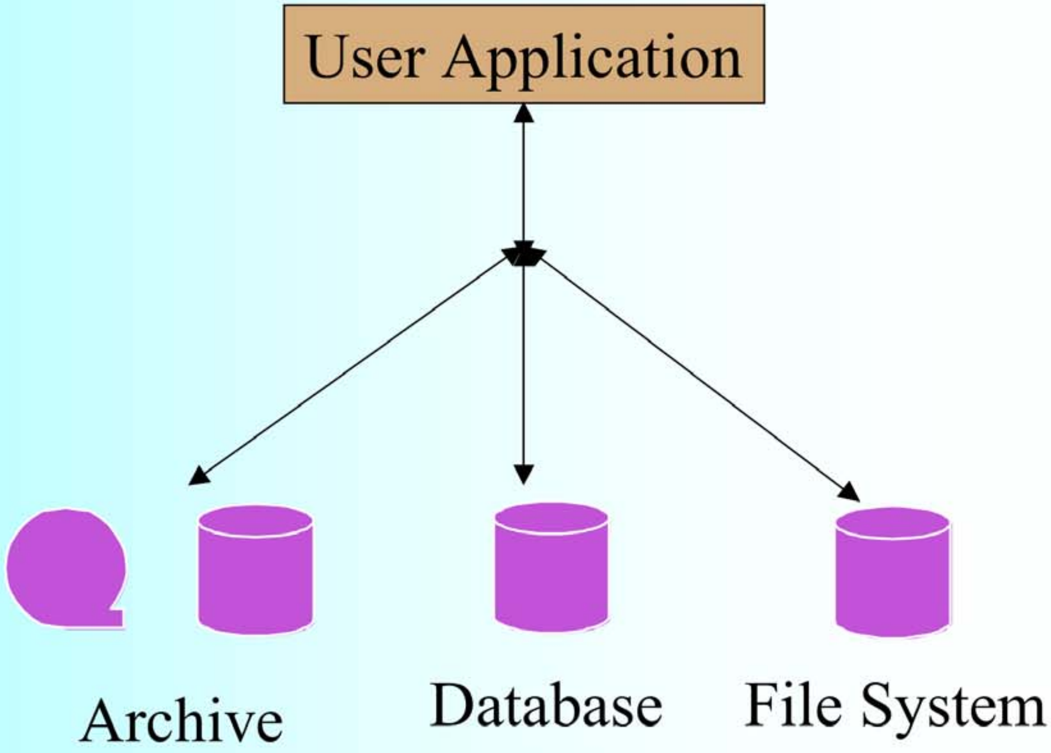
Data Grid Components

- Federated client-server architecture
  - Servers can talk to each other independently of the client
- Infrastructure independent naming
  - Logical names for users, resources, files, applications
- Collective ownership of data
  - Collection-owned data, with infrastructure independent access control lists
- Context management
  - Record state information in a metadata catalog from data grid services such as replication
- Abstractions for dealing with heterogeneity

Data Grid Abstractions

- **Logical name space for files**
  - Global persistent identifier
- **Storage repository virtualization**
  - Standard operations supported on storage systems
- **Information repository virtualization**
  - Standard operations to manage collections in databases
- **Access virtualization**
  - Standard interface to support alternate APIs
- **Latency management mechanisms**
  - Aggregation, parallel I/O, replication, caching
- **Security interoperability**
  - GSSAPI, inter-realm authentication, collection-based authorization

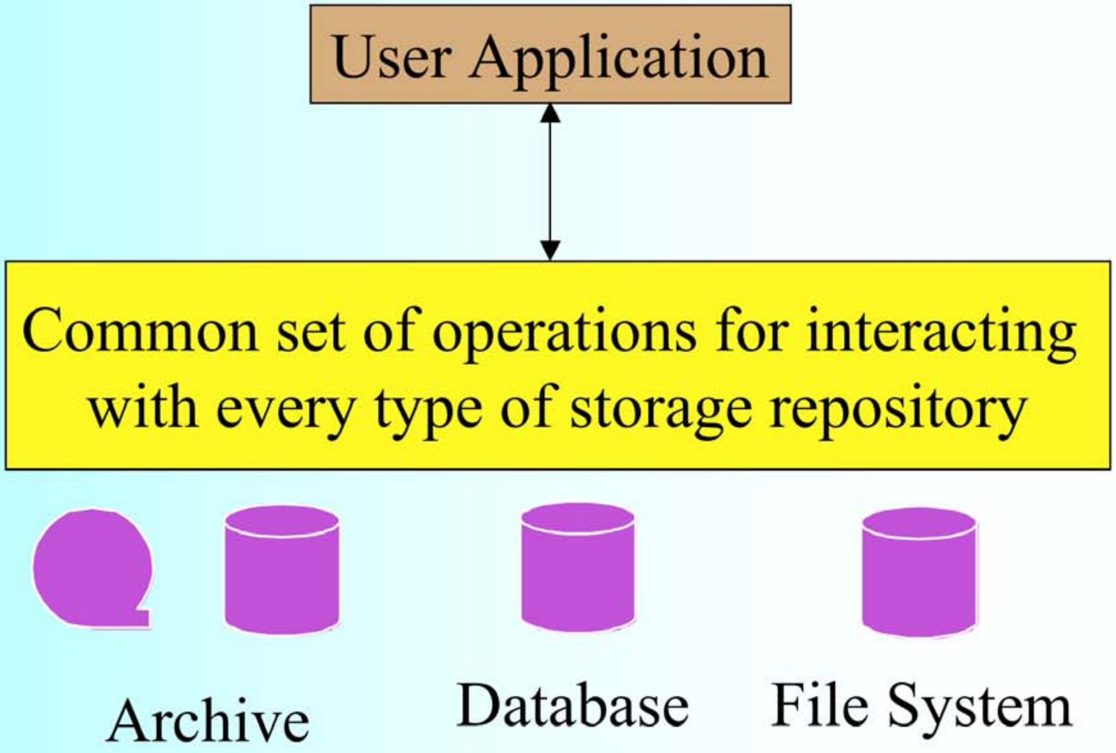
Storage Repository Virtualization



Storage Repository Virtualization

Remote operations

- Unix file system
- Latency management
- Procedures
- Transformations
- Third party transfer
- Filtering
- Queries



User Application

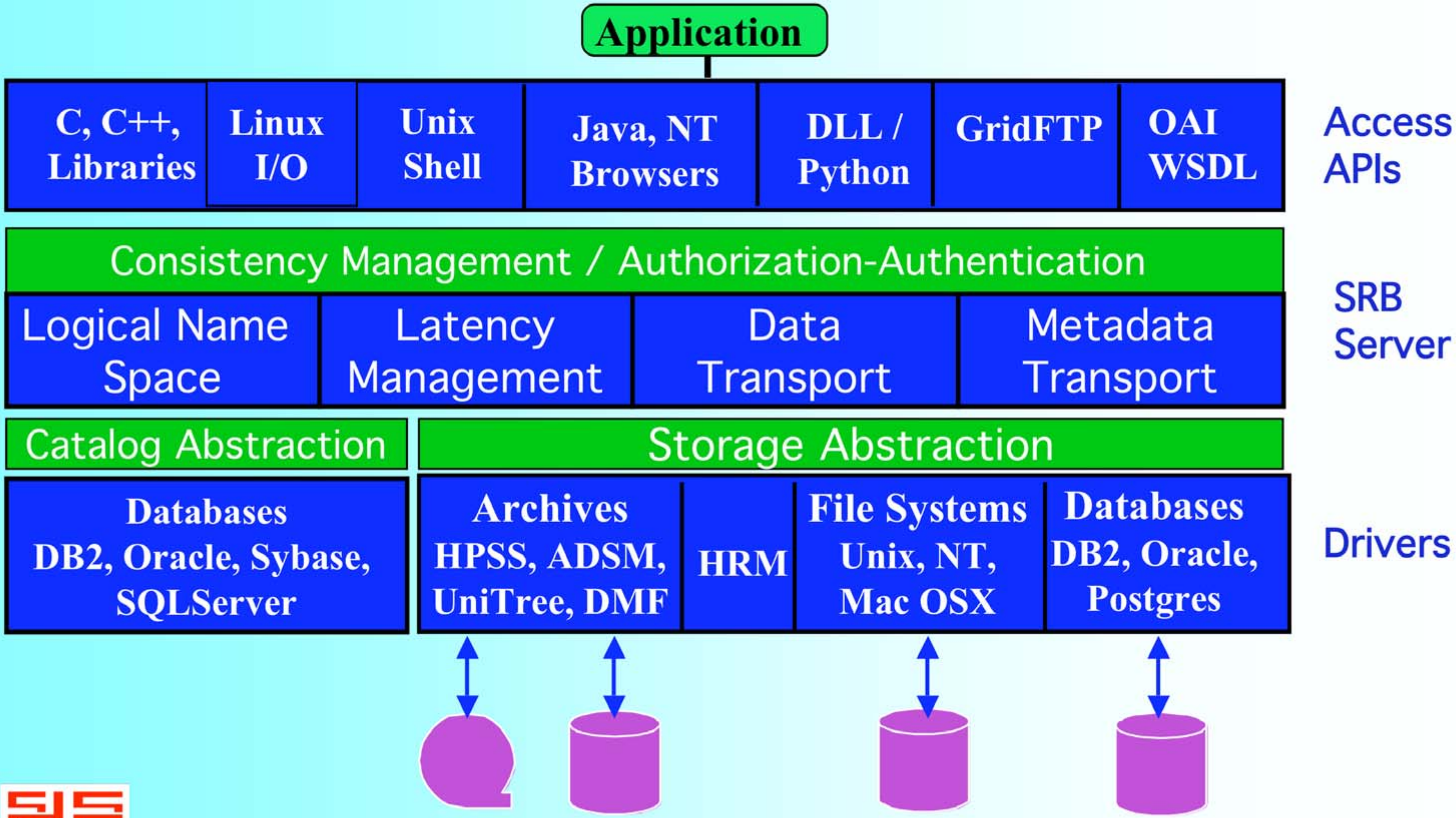
Common set of operations for interacting with every type of storage repository

Archive

Database

File System

SDSC Storage Resource Broker & Meta-data Catalog



Application

C, C++, Libraries | Linux I/O | Unix Shell | Java, NT Browsers | DLL / Python | GridFTP | OAI WSDL — Access APIs

Consistency Management / Authorization-Authentication

Logical Name Space | Latency Management | Data Transport | Metadata Transport — SRB Server

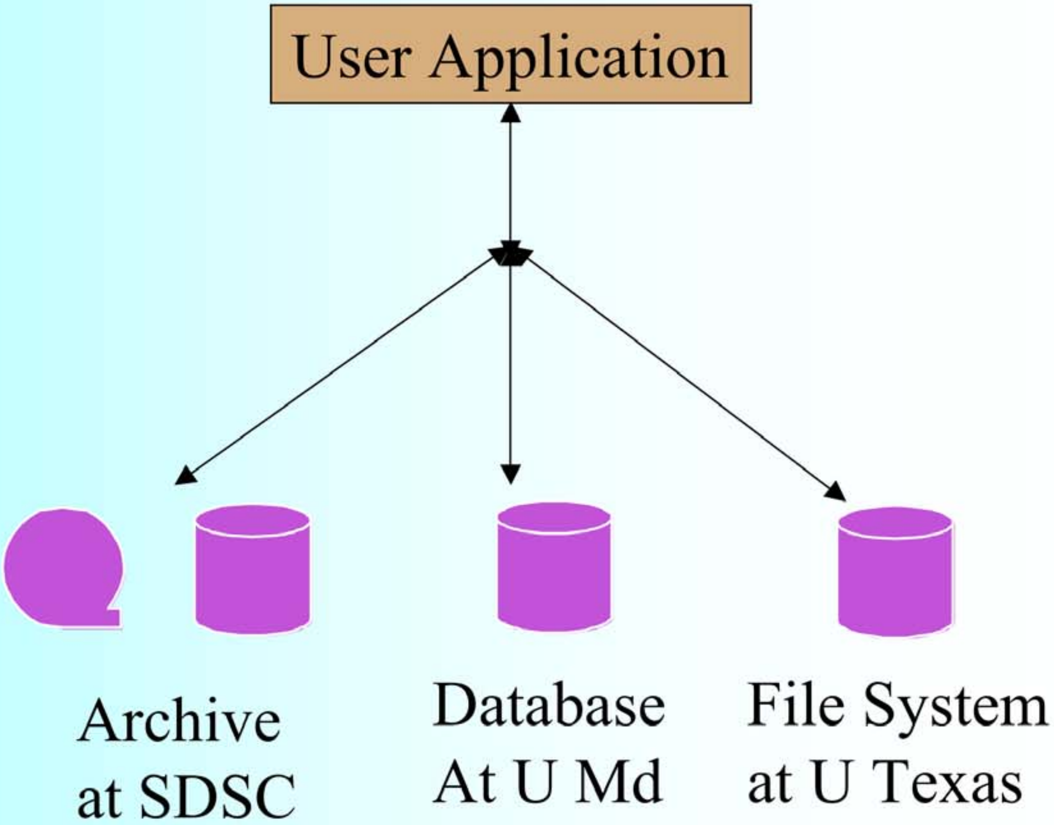
Catalog Abstraction | Storage Abstraction

Databases DB2, Oracle, Sybase, SQLServer | Archives HPSS, ADSM, UniTree, DMF | HRM | File Systems Unix, NT, Mac OSX | Databases DB2, Oracle, Postgres — Drivers

Production Data Grid

- SDSC Storage Resource Broker
  - Federated client-server system, managing
    - Over 90 TBs of data at SDSC
    - Over 16 million files
  - Manages data collections stored in
    - Archives (HPSS, UniTree, ADSM, DMF)
    - Hierarchical Resource Managers
    - Tapes, tape robots
    - File systems (Unix, Linux, Mac OS X, Windows)
    - FTP sites
    - Databases (Oracle, DB2, Postgres, SQLserver, Sybase, Informix)
    - Virtual Object Ring Buffers

Data Virtualization

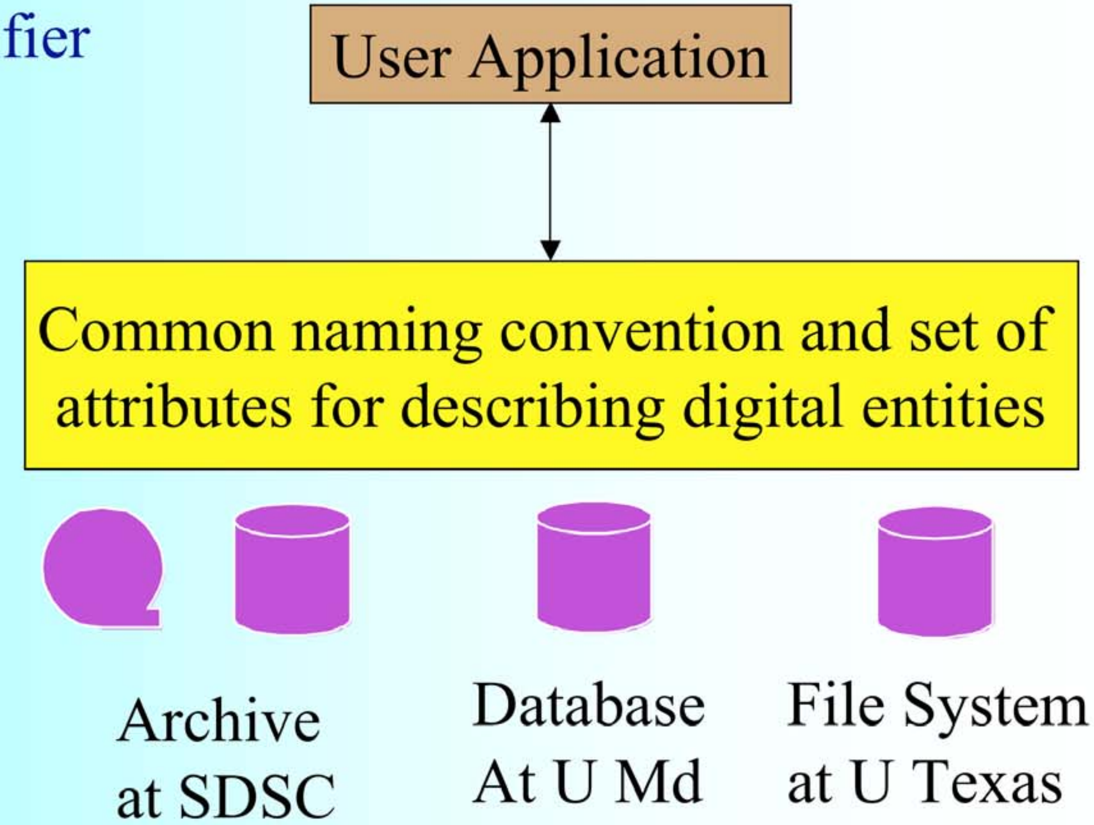


Data Virtualization

Logical name space
- Location independent identifier
- Persistent identifier

Collection owned data
- Access controls
- Audit trails
- Checksums
- Descriptive metadata

Inter-realm authentication
- Single sign-on system


User Application

Common naming convention and set of attributes for describing digital entities

Archive at SDSC

Database At U Md

File System at U Texas

Logical Name Space

- Global, location-independent identifiers for digital entities
  - Organized as collection hierarchy
  - Attributes mapped to logical name space
    - Attributed managed in a database
- Types of administrative metadata
  - Physical location of file
  - Owner, size, creation time, update time
  - Access controls

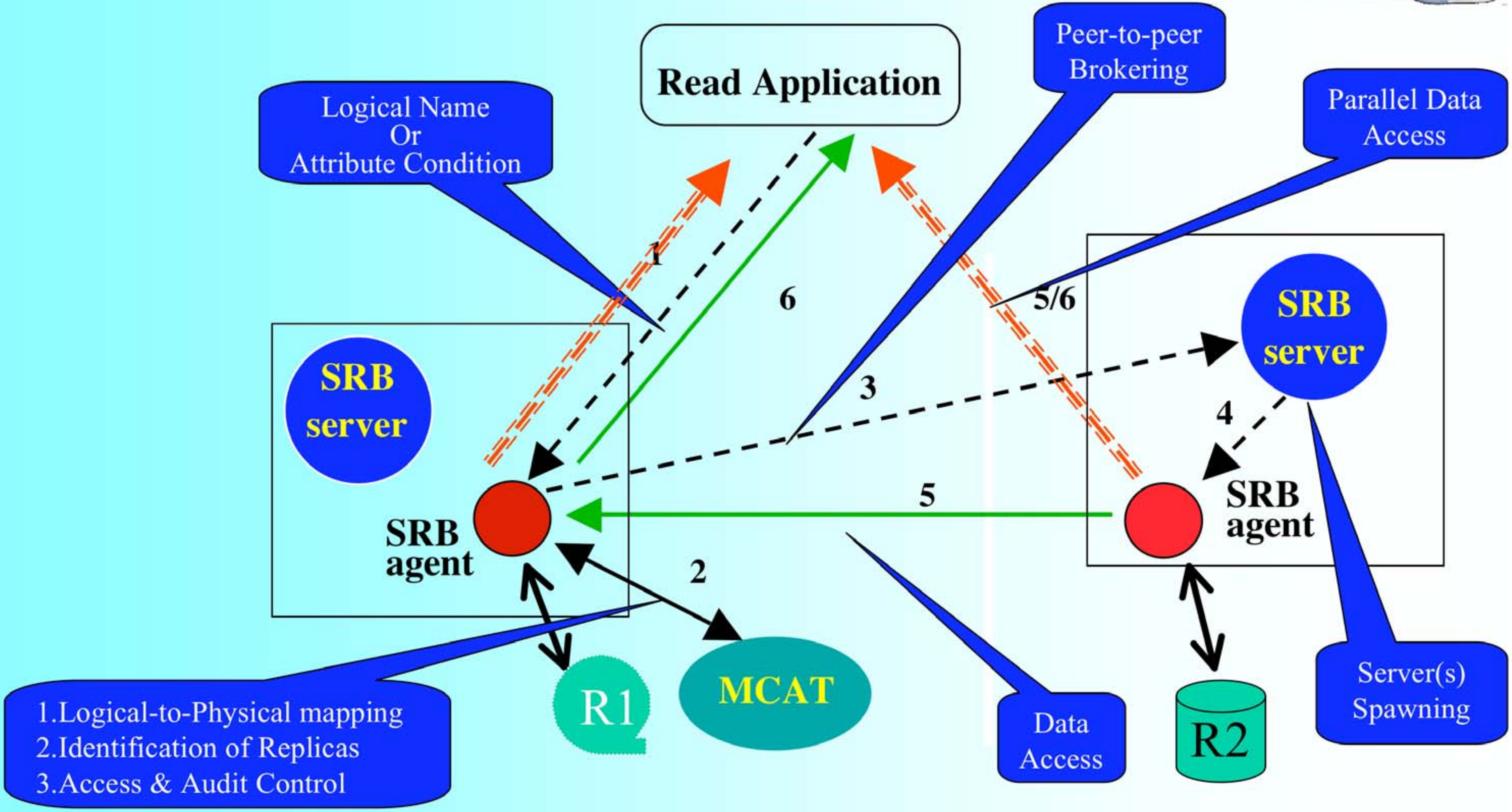
File Identifiers

- Logical file name
  - Infrastructure independent
  - Used to organize files into a collection hierarchy
- Globally unique identifier
  - GUID for asserting equivalence across collections
- Descriptive metadata
  - Support discovery
- Physical file name
  - Location of file

Mappings on Name Space

- Define logical resource name
  - List of physical resources
- Replication
  - Write to logical resource completes when all physical resources have a copy
- Load balancing
  - Write to a logical resource completes when copy exist on next physical resource in the list
- Fault tolerance
  - Write to a logical resource completes when copies exist on “k” of “n” physical resources

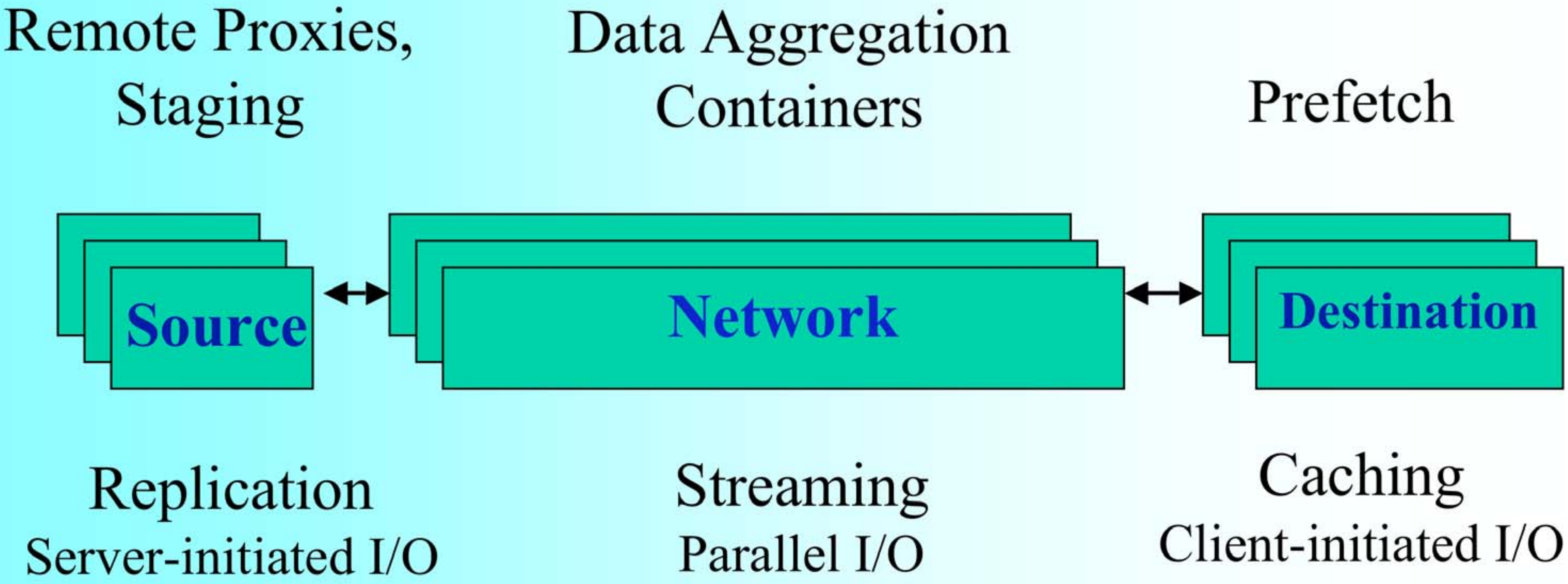
Federated SRB server model



Latency Management - Bulk Operations

- Bulk register
  - Create a logical name for a file
  - Load context (metadata)
- Bulk load
  - Create a copy of the file on a data grid storage repository
- Bulk unload
  - Provide containers to hold small files and pointers to each file location
- Requests for bulk operations for delete, access control, …

SRB Latency Management



Remote Proxies, Staging

Data Aggregation Containers

Prefetch

Source ↔ Network ↔ Destination

Replication

Server-initiated I/O

Streaming

Parallel I/O

Caching

Client-initiated I/O

Remote Proxies

- Extract image cutout from Digital Palomar Sky Survey
  - Image size 1 Gbyte
  - Shipped image to server for extracting cutout took 2-4 minutes (5-10 Mbytes/sec)
- Remote proxy performed cutout directly on storage repository
  - Extracted cutout by partial file reads
  - Image cutouts returned in 1-2 seconds
- Remote proxies are a mechanism to aggregate I/O commands

NASA Data Grids

- NASA Information Power Grid
  - NASA Ames, NASA Goddard
  - Distributed data collection using the SRB
- ESIP federation
  - Led by Joseph JaJa (U Md)
  - Federation of ESIP data resources using the SRB
- NASA Goddard Data Management System
  - Storage repository virtualization (Unix file system, Unitree archive, DMF archive) using the SRB
- NASA EOS Petabyte store
  - Storage repository virtualization for EMC persistent store using the Nirvana version of SRB

Access Abstraction

Example - Data Assimilation Office

HSI has implemented metadata schema in SRB/MCAT

Origin: host, path, owner, uid, gid, perm_mask, [times]

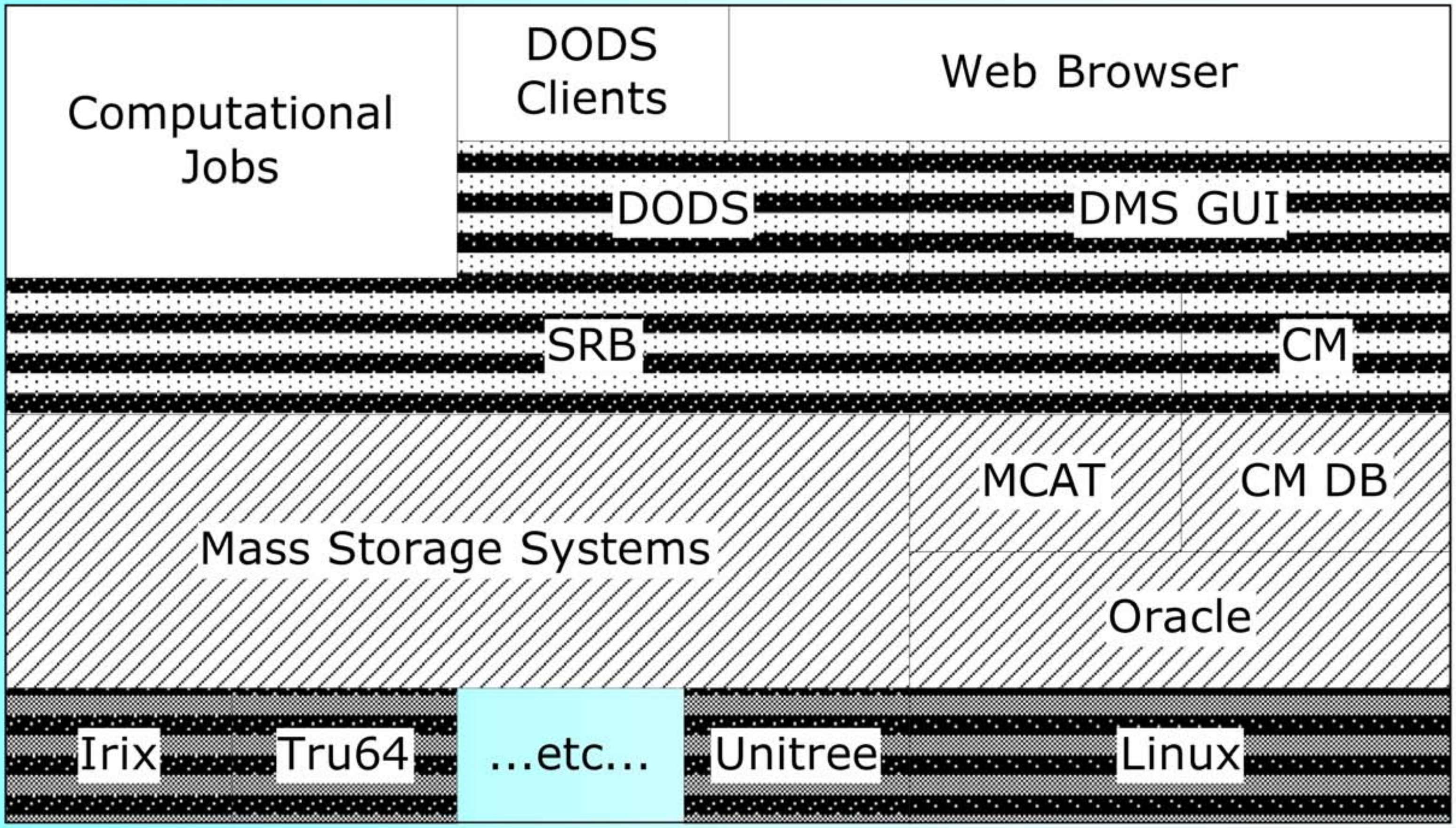
Ingestion: date, user, user_email, comment

Generation: creator (name, uid, user, gid), host (name, arch, OS name & flags), compiler (name, version, flags), library, code (name, version), accounting data

Data description: title, version, discipline, project, language, measurements, keywords, sensor, source, prod. status, temporal/spatial coverage, location, resolution, quality

Fully compatible with GCMD

Data Management System: Software Architecture



Computational Jobs

DODS Clients

Web Browser

DODS

DMS GUI

SRB

CM

Mass Storage Systems

MCAT

CM DB

Oracle

Irix

Tru64

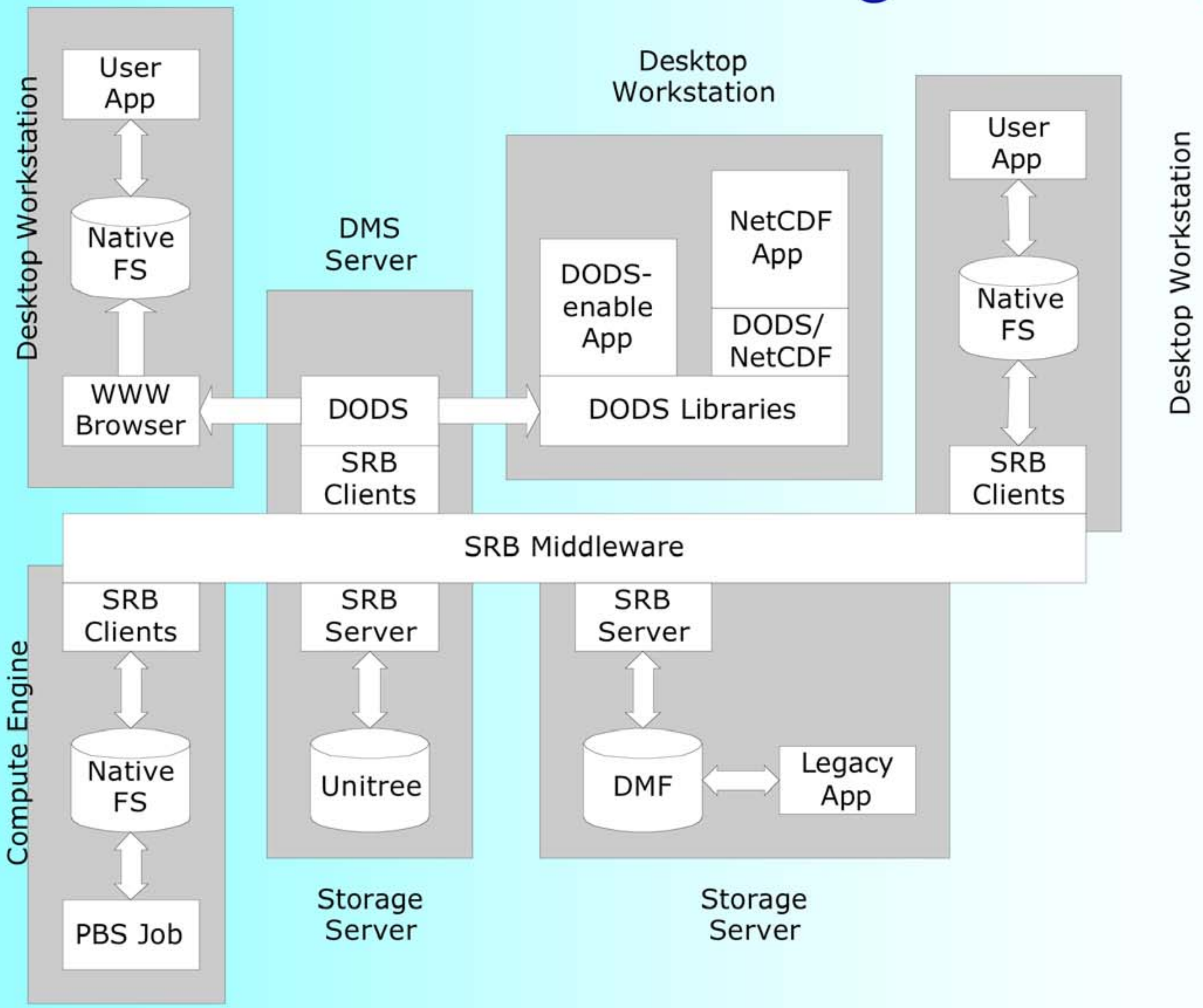
...etc...

Unitree

Linux

HALCYON SYSTEMS, INC.

DODS Access Environment Integration



Zone SRB Federation

- Mechanisms to impose consistency and access constraints when sharing:
  - Resources
    - Controls on which zones may use a resource
  - User names (user-name / domain / SRB-zone)
    - Users may be registered into another domain, but retain their home zone, similar to Shibboleth
  - Data files
    - Controls on who specifies replication of data
  - Context metadata
    - Controls on who manages updates to metadata

Knowledge Based Data Grid Roadmap

	Ingest Services		Management		Access Services
Knowledge	Relationships Between Concepts	XTM DTD	Knowledge Repository for Rules	Rules - KQL	Knowledge or Topic-Based Query / Browse
	(Model-based Access)				
Information	Attributes Semantics	XML DTD	Information Repository	SDLIP	Attribute- based Query
	(Data Handling System)				
Data	Fields Containers Folders	MCAT/HDF	Storage (Replicas, Persistent IDs)	Grids	Feature-based Query

For More Information

Reagan W. Moore
San Diego Supercomputer Center

moore@sdsc.edu

http://www.npaci.edu/DICE

http://www.npaci.edu/DICE/SRB

http://www.npaci.edu/dice/srb/mySRB/mySRB.html

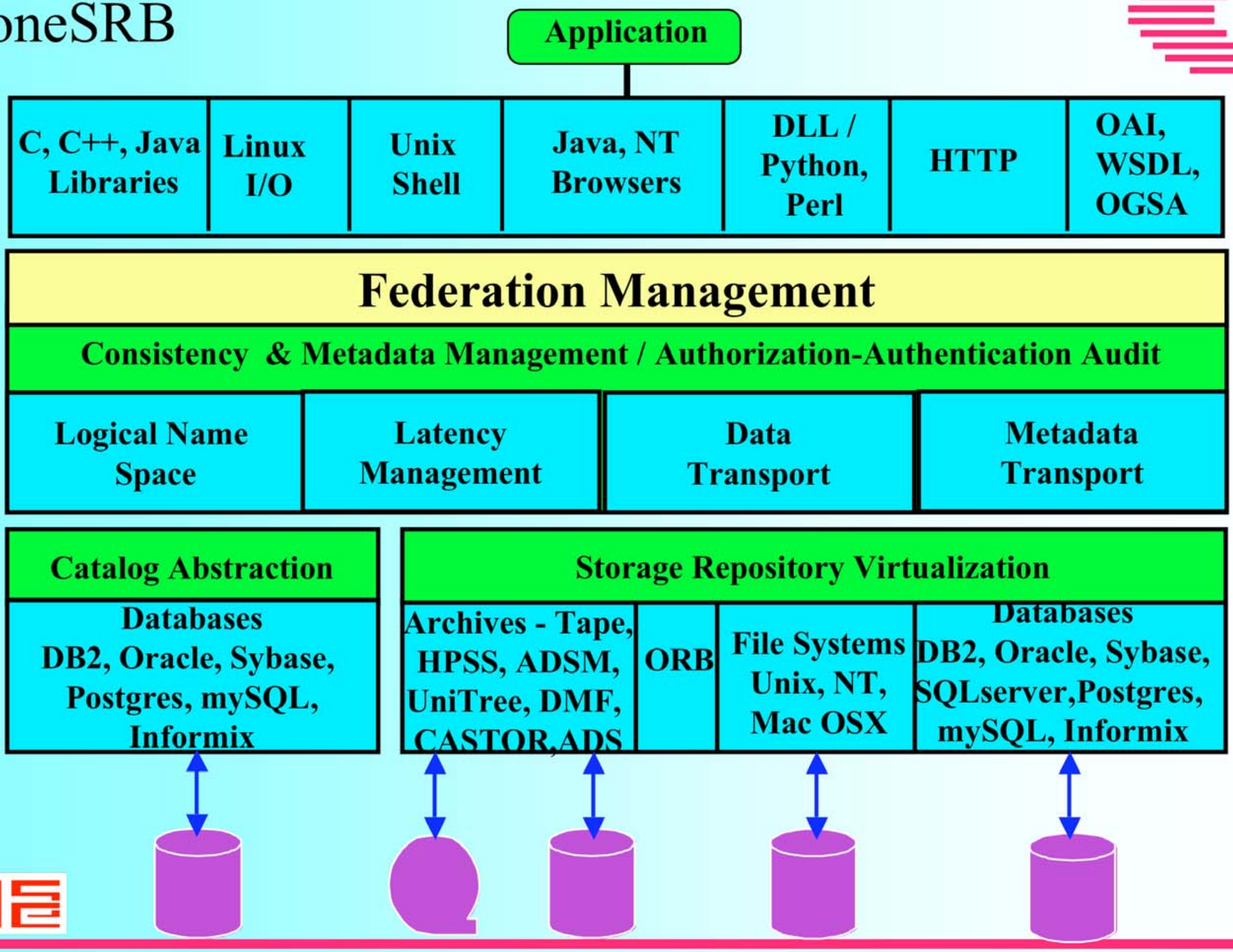
Data Grid Federation

- Data grids provide the ability to name, organize, and manage data on distributed storage resources
- Federation provides a way to control sharing of resources, users, data and metadata between independent data grids.
- We call each data grid a “zone”, hence zoneSRB

Data Grid Federation

- Consistency constraints in federations
- Cross-register a digital entity from one collection into another
    - Who manages the access control lists?
    - Who maintains consistency between context and content?
- How can federation systems be characterized?
    - Peer-to-peer sharing between data grids
    - Hierarchical organization of data grids

Data Grid Federation - zoneSRB



Application

C, C++, Java Libraries | Linux I/O | Unix Shell | Java, NT Browsers | DLL / Python, Perl | HTTP | OAI, WSDL, OGSA

Federation Management

Consistency & Metadata Management / Authorization-Authentication Audit

Logical Name Space | Latency Management | Data Transport | Metadata Transport

Catalog Abstraction

Databases DB2, Oracle, Sybase, Postgres, mySQL, Informix

Storage Repository Virtualization

Archives - Tape, HPSS, ADSM, UniTree, DMF, CASTOR,ADS | ORB | File Systems Unix, NT, Mac OSX | Databases DB2, Oracle, Sybase, SQLserver,Postgres, mySQL, Informix

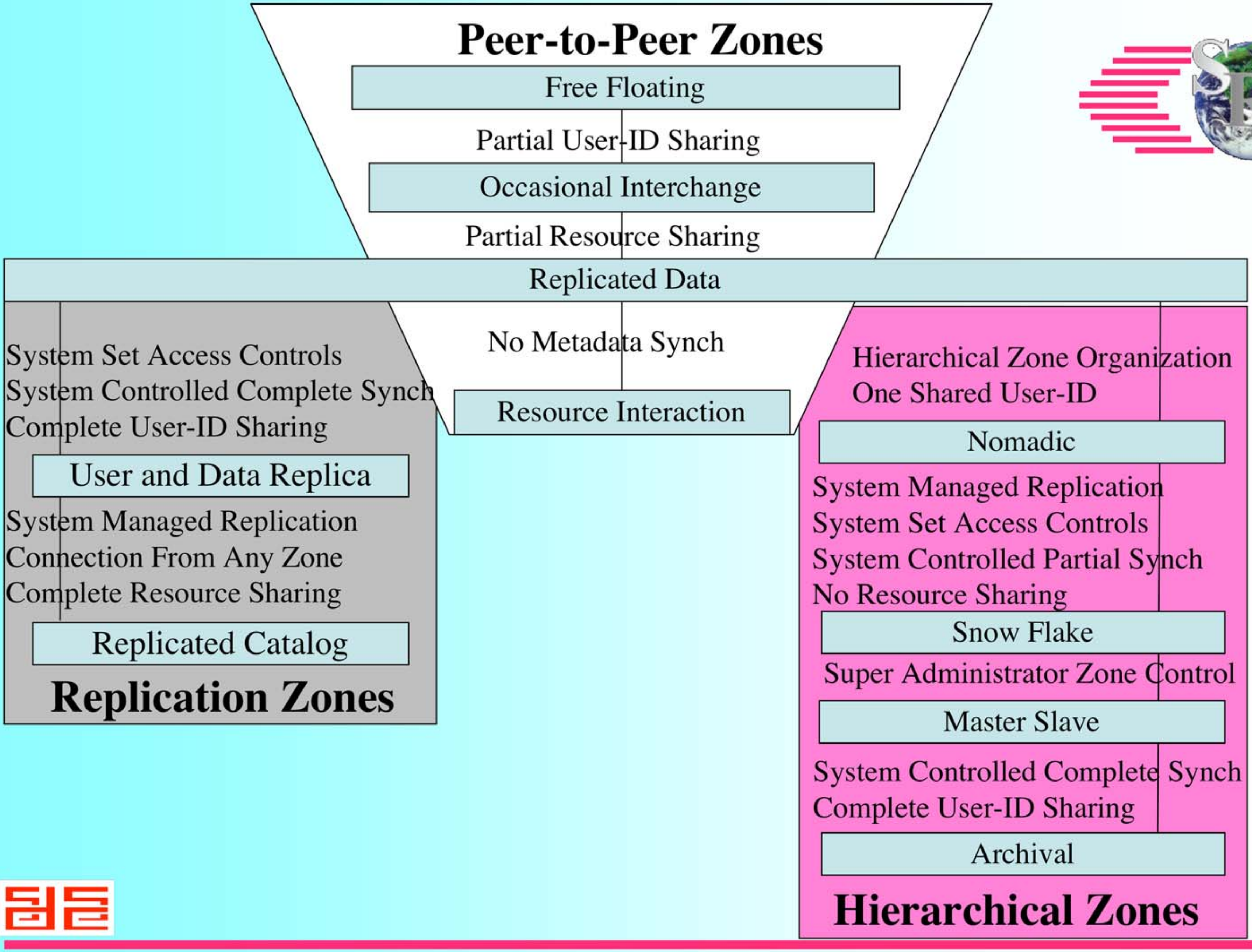
Peer-to-Peer Federation

1. Occasional Interchange – for specified users
2. Replicated Catalogs – entire state information replication
3. Resource Interaction – data replication
4. Replicated Data Zones – no user interactions between zones
5. Master-Slave Zones – slaves replicate data from master zone
6. Snow-Flake Zones – hierarchy of data replication zones
7. User / Data Replica Zones – user access from remote to home zone
8. Nomadic Zones “SRB in a Box” – synchronize local zone to parent zone
9. Free-floating “myZone” – synchronize without a parent zone
10. Archival “BackUp Zone” – synchronize to an archive

SRB Version 3.0.1 released December 19, 2003

Principle peer-to-peer federation approaches
(1536 possible combinations)

Zone SRB	Zone Organization	Zone interaction control	Consistency Management	User Connection Point to access files	Data Access Control Setting	Metadata synchroni-zation	Resource sharing	User-ID sharing between zones
	Zones	Zones	Collections	Files	Files	Metadata	Resources	User names
Free Floating Zones	Peer-to-Peer	Local Admin	User-specified data publication	From home zone	User set access controls	User controlled synchronization	None	None
Occasional Interchange	Peer-to-Peer	Local Admin	User specified	From home zone	User set access controls	User controlled synchronization	None	Partial
Replicated Data Zones	Peer-to-Peer	Local Admin	User-specified replication	From home zone	User set local access controls	User controlled synchronization	Partial	Partial, user establishes own accounts
Resource Interaction	Peer-to-Peer	Local Admin	User-specified replication	From home zone	User set access controls	None	Partial shared resource for replication	Partial
User and Data Replica Zones	Peer-to-Peer	Local Admin	User-specified replication	From home zone	System set access controls	System controlled complete synchronization	Partial	Complete
Replicated Catalog	Peer-to-Peer	Local Admin	System managed name conflict resolution	From any zone	System replicated access controls	System controlled complete synchronization	All zones share resources	Complete
Snow Flake Zones	Hierarchical	Local Admin	System managed replication in hierarchy of zones	From home zone	System set access controls	System controlled partial synchronization	None	One
Master-Slave Zones	Hierarchical	Super Admin	System-managed replication to slave	From home zone	System set access controls	System controlled partial synchronization	None	One
Archival zones	Hierarchical	Super Admin	System-managed versioning to parent zone	From home zone	System set access controls	System controlled complete synchronization	None	Complete
Nomadic Zones	Hierarchical	Local Admin	User-managed replication to parent zone	From home zone	User set access controls	User controlled synchronization	Partial	One



Peer-to-Peer Zones

Free Floating

Partial User-ID Sharing

Occasional Interchange

Partial Resource Sharing

Replicated Data

No Metadata Synch

Resource Interaction

System Set Access Controls
System Controlled Complete Synch
Complete User-ID Sharing

User and Data Replica

System Managed Replication
Connection From Any Zone
Complete Resource Sharing

Replicated Catalog

Replication Zones

Hierarchical Zone Organization
One Shared User-ID

Nomadic

System Managed Replication
System Set Access Controls
System Controlled Partial Synch
No Resource Sharing

Snow Flake

Super Administrator Zone Control

Master Slave

System Controlled Complete Synch
Complete User-ID Sharing

Archival

Hierarchical Zones

Data Grid Demonstration

- Use web browser to access a collection housed at SDSC
- Retrieve an image
- Browse through a collection
- Search for a file
- Examine grid federation

Grid Bricks

- Integrate data management system, data processing system, and data storage system into a modular unit
  - Commodity based disk systems (1 TB)
  - Memory (1 GB)
  - CPU (1.7 Ghz)
  - Network connection (Gig-E)
  - Linux operating system
- Data Grid technology to manage name spaces
  - User names (authentication, authorization)
  - File names
  - Collection hierarchy

Data Grid Brick

- Hardware components
  - Intel Celeron 1.7 GHz CPU
  - SuperMicro P4SGA PCI Local bus ATX mainboard
  - 1 GB memory (266 MHz DDR DRAM)
  - 3Ware Escalade 7500-12 port PCI bus IDE RAID
  - 10 Western Digital Caviar 200-GB IDE disk drives
  - 3Com Etherlink 3C996B-T PCI bus 1000Base-T
  - Redstone RMC-4F2-7 4U ten bay ATX chassis
  - Linux operating system
- Cost is $2,200 per Tbyte plus tax
- Gig-E network switch costs $500 per brick
- Effective cost is about $2,700 per TByte

Grid Bricks at SDSC

- Used to implement “picking” environments for 10-TB collections
  - Web-based access
  - Web services (WSDL/SOAP) for data subsetting
- Implemented 15-TBs of storage
  - Astronomy sky surveys, NARA prototype persistent archive, NSDL web crawls
- Must still apply Linux security patches to each Grid Brick
- Grid bricks managed through SRB
  - Logical name space, User Ids, access controls
  - Load leveling of files across bricks

SDSC SRB Team

- Reagan Moore
- Michael Wan
- Arcot Rajasekar
- Wayne Schroeder
- Arun Jagatheesan
- Charlie Cowart
- Lucas Gilbert
- George Kremenek
- Sheau-Yen Chen
- Bing Zhu
- Roman Olschanowsky (BIRN)
- Vicky Rowley (BIRN)
- Marcio Faerman (SCEC)
- Antoine De Torcy (IN2P3)
- Students & *emeritus*
  - Erik Vandekieft
  - Reena Mathew
  - Xi (Cynthia) Sheng
  - Allen Ding
  - Grace Lin
  - Qiao Xin
  - Daniel Moore
  - Ethan Chen
  - Jon Weinburg